Logic-based Remodeling of the Digital Anatomist Foundational Model

Rainer Beck Stefan Schulz
Freiburg University Hospital, Department of Medical Informatics (`http://www.imbi.uni-freiburg.de/medinf`)

Abstract

This paper describes a development cycle for the engineering of large knowledge bases: A graphical tool is used for editing and the content is transformed into a logic-based representation language. This representation is used to check the consistency of the knowledge base as well as to facilitate the reviewing process. Showing the usefulness of this approach, aspects of the Digital Anatomist Foundational Model will be transformed into a Description Logics representation. We introduce a special modeling technique to account for the representation of the complex part/whole relationships in the biomedical domain.

INTRODUCTION

As more and more facts are gathered in the domain of life sciences, so do biomedical concept systems grow. The emphasis in most of these systems is on the comprehensiveness of coverage; only a few formalize concept representation. The UMLS [15] is a specific example since it merges vocabularies which differ in terms of conceptualization, resulting in poorly defined semantics. This deficiency can be largely attributed to the different contexts of its source vocabularies. However, there now is an increased trend towards strict semantics, resulting in concept systems capable of supporting formal reasoning [8, 14].

Physiology, pathology, molecular biology and other subdomains are characterizable by the change they inflict on underlying physical structure. Therefore formal representation of the structural sciences (like anatomy, structural biology) is particularly important. So it is to be expected that an elaborated, common structural model would facilitate the conceptualization and representation of all biomedical domains. Moreover this constitutes a critical requirement for a principled representation of knowledge about diagnostic and therapeutic procedures.

As a representation of anatomical structure, the Digital Anatomist Foundational Model (FM) [11] is outstanding in respect to coverage and formal strictness: It describes canonical anatomical knowledge using rather precise semantics. The design of the model is guided by the very tight formal Aristotelian principles of genus and differentiae [5] (most prominently implemented in the upper level ontology). Taxonomic and partonomic hierarchies are strictly separated. Because of its coverage (more than 67,000 concepts) and its design principles, the FM presents an ideal candidate to be represented in a logic-based language.

Whereas the taxonomic structure of the FM may be already interpreted in terms of a formal ontology, the semantics of the relations *part-of* and *has-part* needs to be further refined. This is particularly desirable since various fundamental properties of biological entities (like *contains*, *has-function*) interact with part / whole relationships between concepts. Therefore, the expressive representation of partonomy is of paramount importance. As a consequence of the comprehensiveness of the FM, ameliorating the part/whole relationships proves to be a complex and challenging task: Small changes of the partonomy may escalate throughout the hierarchy and lead to arbitrarily complex results.

In this paper, we propose a development cycle suitable for the engineering of large knowledge bases obeying our principles of “single point of edit” and “minimum manual intervention”. Furthermore, we apply this method to the Foundational Model and show how to transform its taxonomy and partonomy into a representation language with formal semantics based on description logics (DL) [1]. We use a distinctive modeling pattern to address the specific needs for representing part/whole relations.

DEVELOPMENT CYCLE

The engineering of large knowledge bases on code level can be a daunting if not impossible task. To support the knowledge engineer in the process, we propose the following cycle (cf. Figure 1):

1. Concept Creation or Change
   In order to effect all modifications in a precise manner, the knowledge engineer must be able to identify the context he or she is working in: Which concepts are defined? Which relations hold between them? This process can be facilitated if there exists a comprehensive graphical representation including search and navigation facilities sufficient for the task. However, significant advantage will be lost if parts of the editing are left to be made outside this

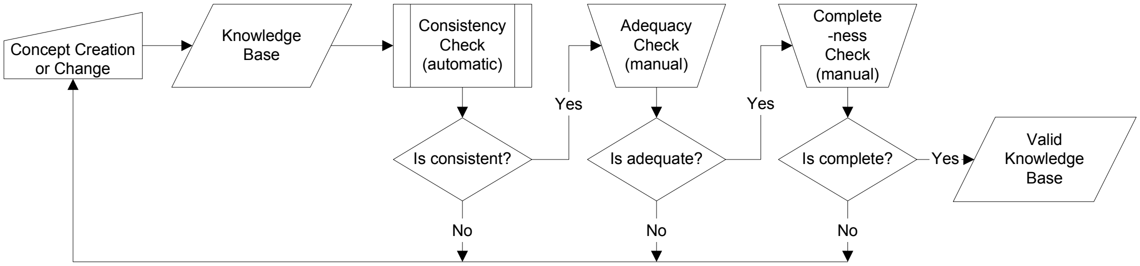

Figure 1: Development Cycle

environment because of lack of support. The "single point of edit" provided by the environment should not be compromised. We consider the kind of environment described a necessity for working with large knowledge bases.

2. Consistency Check
   After changing or adding to the knowledge represented, consistency checks should be performed: Semantic integrity is easily lost in large knowledge bases, especially when role propagation [1] along several axes is involved. Therefore, we propose at least semi-automatical checks by an inference engine which can provide a status report. Thus, inconsistencies as well as terminological cycles will be identified and can be corrected by iterating the first step. It is also highly desirable to integrate these mechanisms into the editing environment for automatic consistency checks.

3. Adequacy Check
   For good reasons, implicit knowledge as well as complete role propagation are usually not fully accessible in a knowledge editing tool. As these inferences are prerequisites for checking the adequacy of the representation entered, they have to be made accessible to the knowledge engineer: Any inference engine used should provide a suitable display interface to allow the analysis of all concepts and the accuracy of their representation by the knowledge engineer. Should the representation be incomplete or inadequate, steps one and two might be repeated until the results are satisfactory.

4. Completeness Check
   The criteria of adequacy being satisfied, the check for completeness may take place in which the knowledge engineer compares the coverage in the representation with the domain to be described. The display and browsing tools required for checking adequacy may be used for this purpose. Eventually, the process results in a valid and comprehensive knowledge base representing the domain of interest.

If the transition between the steps described is well automated, it will be possible to handle large and very large knowledge bases with "minimum manual intervention". The danger of losing track of the modifications can be minimized and their respective effects on the representation can be controlled at the same time. The merging of all steps into an editing tool would be the ultimate solution. However, considering the complexity involved with reasoning, online processing and feedback seems rather unlikely for the time being.

PART / WHOLE REASONING

Although the FM represents the relations *part-of* and *has-part* (called *part* in the FM) between its concepts, it does not define the semantics of these relations formally. We therefore propose the following axioms in which *A* and *B* are concepts, and *r* standing for the relation *part-of* or *has-part*:

- $A\ r\ B$ will be interpreted as $A \sqsubseteq \exists r.B$ [2]
- r is transitive. Hence, $A \sqsubseteq \exists r.B$ and $B \sqsubseteq \exists r.C$ imply $A \sqsubseteq \exists r.C$
- r is irreflexive and antisymmetric.
- Self references are disallowed: $A \sqsubseteq B \sqcap \exists r.B$ is not a valid expression.

[1] Inheritance of roles along non-taxonomic axes [2]

[2] The DL constructors used in this article are a subset of the standard DL language $\mathcal{ALC}$:

– Concepts and relations
– Taxonomic subsumption ($\sqsubseteq$), e.g. $Woman \sqsubseteq Human$
– Full existential quantification ($\exists$), e.g. $\exists has\text{-}part.Nose$ comprises all individuals which have at least 1 nose.
– Intersection ($\sqcap$), e.g. $Woman \sqcap \exists has\text{-}child.Woman$ denotes all individuals which satisfy both of the criteria (in this case a woman with at least 1 daughter).
– Union ($\sqcup$), e.g. $male \sqcup female$ denotes all individuals which satisfy at least one of the criteria (male or female)

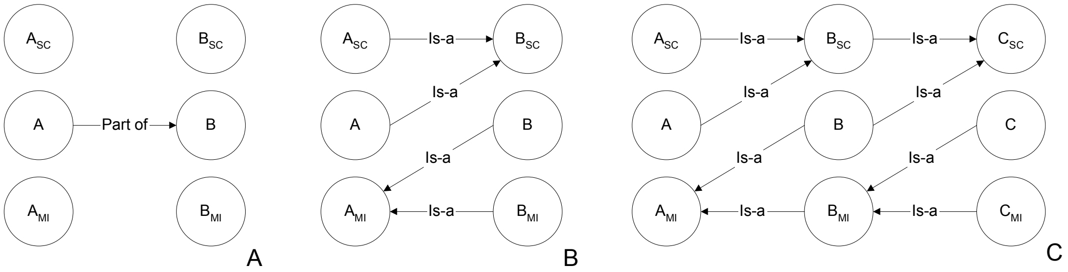

Figure 2: Transformation of partitive information into taxonomic hierarchies using ePI (*SI* and *MC* nodes are not shown for sake of clarity). A depicts the relation $A \sqsubseteq \exists part\text{-}of B$, B this expression in a taxonomic representation and C how transitivity can be modeled by the use of ePI

When looking more closely at part / whole relationships, concepts may have different roles when participating in *part-of/has-part* relations:

- $Appendix \sqsubseteq \exists part\text{-}of.Midgut$
  There exists a *midgut* for every *appendix* and the *appendix* is part of it – therefore, the *midgut* is a **mandatory includer** of an *appendix* as it can not be found without a *midgut* it is part of. Changing the view, the *appendix* is a **specific component** of the *midgut*: If there is an *appendix*, there is a *midgut*.

- $Blood \sqsubseteq \exists has\text{-}part.Erythrocyte$
  Blood always includes *erythrocytes* – *erythrocytes* are a **mandatory constituent** of *blood* as there is no *blood* without *erythrocytes*. Vice versa, *blood* is a **specific includer** of *erythrocytes*: If there is *blood*, there are always *erythrocytes*.

The roles occur in mutually dependent concepts as well (e.g. *CellNucleus* and *NuclearMembrane*).
We introduce the following concepts (A_{sc}, A_{mc}, A_{si}, A_{mi}) as reificators:

- A_{sc}, subsumer of the **specific components** of *A*, comprises all concepts whose instances are related by *part-of* to an instance of *A*.

- A_{mc}, subsumer of the **mandatory constituents** of *A*, comprises all concepts whose instances an instance of *A* is related to by *has-part*.

- A_{si}, subsumer of the **specific includers** of *A*, comprises all concepts whose instances are related by *has-part* to an instance of *A*.

- A_{mi}, subsumer of the **mandatory includers** of *A*, comprises all concepts whose instances an instance of *A* is related to by *part-of*.

This model, called *ePI* (extended part/include) is an extension of the SEP triplet [2] and the PI encoding scheme [12].

One of the uses of this representation pattern is the transformation of partitive relationships into simpler taxonomic ones (cf Fig. 2):

$$BetaCell \sqsubseteq \exists part\text{-}of.Pancreas \quad \text{(2A)}$$

is expressed by the following equivalent statements:

$$BetaCell \sqsubseteq Pancreas_{sc}$$
$$Pancreas \sqsubseteq BetaCell_{mi} \quad \text{(2B)}$$

By introducing

$$BetaCell_{sc} \sqsubseteq Pancreas_{sc}$$
$$Pancreas_{mi} \sqsubseteq BetaCell_{mi} \quad \text{(2B)}$$

specific components of *BetaCell* are also classified as specific components of *Pancreas*, as well as mandatory includers of *Pancreas* are mandatory includers of *BetaCell* (Fig. 2C). This may serve to model the transitivity of *part-of* / *has-part* in DL dialects which do not support transitivity natively.
Furthermore, other relationships can use these reificators to propagate along the part / whole axis (non-anatomical definitions):

$$Alveolus \sqsubseteq \exists location\text{-}of.GasExchange$$
$$Alveolus_{mi} \sqsubseteq \exists location\text{-}of.GasExchange$$
$$Alveolus \sqsubseteq \exists part\text{-}of.PrimaryPulmonaryLobule$$
$$PrimaryPulmonaryLobule \sqsubseteq \exists part\text{-}of.LobeOfLung$$

allows the deduction

$$LobeOfLung \sqsubseteq \exists location\text{-}of.GasExchange$$[3]

Using the modeling approach above, we are able to account for the intricacies involved in the modeling of partonomies in the biomedical domain at the costs of four extra concepts per entity. Furthermore, it allows one to grasp sets of concepts fulfilling restrictions like ”What has-part DNA?”, ”What parts does the arm have?” or ”Where may lymph follicles be found?” by using the appropriate reificators. This is particularly useful for the review of adequacy.

[3] Further elaboration of these properties is beyond the scope of this paper.

KNOWLEDGE EDITING / IMPORT

We will now describe which tools are used in the implementation of this development cycle:

Protégé-2000[4] [6], a frame-based, modular, extensible knowledge-editor provides a graphical user interface that supports the display of concept hierarchies in a tree-like structure (as well as search and navigation facilities) and allows for inheritance of slots and value restrictions. It is used to build as well as modify the FM during the development cycle and provides the "single point of edit" proposed.

The LOOM[5] knowledge representation system [4] is used as a Description Logics implementation because of its maturity and its known capability for handling large scale knowledge bases.

ONTOSAURUS[6] presents the classified concepts in the form of a web based interface to ease the evaluation of adequacy and consistency. Tightly integrated into LOOM, ONTOSAURUS shows a good performance with large amounts of concepts.

To bridge the gap between Protégé and LOOM, we developed a tool which allows to convert the knowledge entered into DL-formalisms while restricting the need of manual intervention to a minimum. It provides means to introduce the eIP pattern as well as filter mechanisms for simple consistency checks. Reading concept definitions from Protégé, we are able to generate code in several terminological languages such as LOOM, FaCT [3] or interchange formats like OWL[7] and DIG[8].

VALIDATION / EVALUATION

Our work is based on a version of the FM dated 16.09.2002. We implemented the development cycle in order to check the consistency and conceptual integrity; hereby focussing mainly on the partonomy.

Taxonomic integrity (*Is-a)*:

After transformation of 61,699 concepts and their sub-/superclass relations into LOOM-code and subjecting them to the classifier, no taxonomic cycles or inconsistencies were found.

Taxonomic / Partonomic Integrity:

After modification of the transformation process to include the *part-of* and *has-part* relations and mapping them to the nodes of the ePI-model, about 280 cyclic definitions were detected. In general, these cycles fall into two categories: First-order and higher-order terminological cycles. First-order cycles are formed by one or two concepts, and the relations *Is-a*, *part-of* and/or *has-part*. 240 cycles belong to this category. Higher-order cycles involve more than two concepts. They may come about accidentally or suggest interesting partitive modeling challenges. Among the cycles found (only the cycle-forming relations given):

1. $PiaMaterOfCerebralHemisphere \sqsubseteq$
   $\exists part\text{-}of.PiaMaterOfCerebralHemisphere$

2. $BoneOfRadius_{si} \sqsubseteq CompactBoneOfRadius_{si}$
   $CompactBoneOfRadius_{si} \sqsubseteq BoneOfRadius_{si}$

3. $SubdivisionOfCorpusCallosum_{mi} \sqsubseteq$
   $IntercerebralCommissure_{mi} \sqsubseteq$
   $CorpusCallosum_{mi} \sqsubseteq$
   $PosteriorForcepsOfCorpusCallosum_{mi} \sqsubseteq$
   $SubdivisionOfCorpusCallosum_{mi}$

Example number one shows a concept which contains a self reference and thus is rejected by the classifier. Number two exemplifies the difficulties in modeling partonomic hierarchies: The concept *BoneOfRadius* seems to comprise the entity "material which the radius is made of" and the entity "bone of radius". Consequently, it has to be split into two different concepts to accommodate the two entities, or one of the relations has to be removed. Number three is a showcase for the problems arising in modeling large knowledge bases - namely loosing track of the extensive interrelationships. The largest taxonomical cycle found comprised 10 concepts. Every cycle found may be attributed to conditions similar to these: It seems obvious that they are not caused by lax modeling principles (in fact, every cycle violated the modelling principles of the FM), but are unavoidable when modelling very large knowledge bases. Once located, the domain experts of the Digital Anatomist group were able to correct these in a very short time span.

DISCUSSION AND CONCLUSIONS

Several investigators have described the challenges involved with the transformation of concept representations into a semantically rigid form: Pisanelli et al. surveyed parts of the UMLS metathesaurus [7] and represented them in DL, without implementing reasoning mechanisms to account for partitive reasoning.

The importance of well-defined semantics and the effects of transition from one system to another were demonstrated in a cross-validation study involving GALEN and the Read Thesaurus [10]. Although the DL dialect involved (GRAIL) supports special formalisms for reasoning with partonomies, the Read thesaurus provided generic hierarchical information only.

[4] http://protege.stanford.edu

[5] http://www.isi.edu/isd/LOOM/

[6] http://www.isi.edu/isd/ontosaurus.html

[7] http://www.w3.org/2001/sw/WebOnt/

[8] http://dl.kr.org/dig/

Therefore the expressiveness of GRAIL could not be fully leveraged.

Schulz et al. [13] showed the necessity of a high degree of manual intervention on code level, elaborating on the transformation of UMLS concepts in the domains of pathology and anatomy to a logic based representation including some partitive reasoning. This modeling approach included the representation of part / whole relations without accounting for the different existential assumptions related to them.

Our approach tries to combine a comprehensive representation of part / whole relationships with good support for the knowledge engineer (minimizing the need for manual intervention on code level). The representation is able to express partonomies in a semantically precise way, while using only a (relatively) inexpressive dialect of DL. The cost is the proliferation of partially redundant concepts.

As the demands on biomedical ontologies grow, we believe that a stable structural foundation becomes more and more necessary if not indispensable. Our analysis suggests that the FM is a good candidate to meet these requirements. To increase its expressivity with respect to partonomic reasoning, we propose to reevaluate the *part-of* and *include* relations: In most cases, where $A \sqsubseteq \exists \mathit{part\text{-}of}.B$ there is $B \sqsubseteq \exists \mathit{has\text{-}part}.A$ to be found as well. As the FM represents canonical anatomy, one may argue that almost every partitive assertion is mandatory for parts and their respective wholes. A body always has teeth and an appendix as its parts. However, a "clinical anatomy ontology" must be able to account for the absence of both. The inclusion of existential assumptions would in turn benefit the detailed modeling of physiological and other processes using role propagation [9].

For a knowledge base of this size, the FM contains a surprisingly small amount of terminological inconsistencies. In addition to its well-defined conceptualization rules, its consistency facilitates the transition to a logic based representation.

Protégé proved to be a viable tool for creating and editing large knowledge bases with a high gain of clarity in comparison with other methods of knowledge editing. Nevertheless, there are still some performance issues when working with very large knowledge bases.

The development cycle - incorporating the principle of "single point of edit" - proved to be a viable way to refine the FM with "minimal manual intervention", alleviating the need of time consuming and error prone code-level editing. In addition, the work reported in this paper represents the first step towards the comprehensive formal representation of the FM, which may serve as a prototype for similar construction of other large (biomedical) knowledge bases.

ACKNOWLEDGEMENTS

I thank C. Rosse and his research group at the University of Washington for providing the FM and many fruitful discussions.

References

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider. *The Description Logic Handbook*. Cambridge University Press, 2003.

[2] U. Hahn, S. Schulz, and M. Romacker. Partonomic reasoning as taxonomic reasoning in medicine. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 271–276, 1999.

[3] Ian R. Horrocks. Using an expressive description logic: FaCT or fiction? In *KR'98: Principles of Knowledge Representation and Reasoning*, pages 636–645, 1998.

[4] R. M. MacGregor. A description classifier for the predicate calculus. In *Proceedings of the 12th National Conference on Artificial Intelligence*, volume 1, pages 213–220, 1994.

[5] J. Michael, J.L.V. Meijino, and C. Rosse. The role of definitions in biomedical concept representation. In *AMIA 2001*, pages 463–467, 2001.

[6] N. F. Noy, R. W. Fergerson, and M. A. Musen. The knowledge model of protégé-2000. In *2th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*.

[7] Domenico M. Pisanelli, Aldo Gangemi, and Geri Steve. An ontological analysis of the UMLS metathesaurus. In *AMIA'98 – Proceedings of the 1998 AMIA Annual Fall Symposium.*, pages 810–814, 1998.

[8] A. Rector, A. Gangemi, Galeazzi E., A. J. Glowinski, and A. Rossi-Mori. The GALEN model schemata for anatomy. In *MIE 94 – Medical Informatics Europe*, pages 229–233, 1994.

[9] A. L. Rector. Analysis of propagation along transitive roles. In *DL02 - International Workshop on Description Logics*, 2002. Published as CEUR Workshop Proceedings (CEUR-WS.org) via http://CEUR-WS.org/Vol -53/.

[10] Jeremy E. Rogers, Colin Price, Alan Rector, W. Daniel Solomon, and Nick Smeijko. Validating clinical terminology structures: Integration and cross-validation of READ THESAURUS and GALEN. In *AMIA'98 – Proceedings of the 1998 AMIA Annual Fall Symposium*, pages 845–849, 1998.

[11] C. Rosse and J.L.V. Mejino. A reference ontology for bioinformatics: The foundational model of anatomy. *JBI in press*, 2003.

[12] S. Schulz. Bidirectional mereological reasoning in anatomical knowledge bases. In *AMIA 2001*, pages 607–611.

[13] S. Schulz and U. Hahn. Medical knowledge reengineering – converting major portions of the UMLS into a terminological knowledge base. *International Journal of Medical Informatics*, 64:207–221, 2001.

[14] K. A. Spackman. Normal forms for description logic expression of clinical concepts in SNOMED RT. In *AMIA 2001*, pages 627–631.

[15] UMLS. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2002.